

**NEURAL RESPONSE BASED SPEAKER IDENTIFICATION
UNDER NOISY CONDITION**

LEYLA ROOHISEFAT

**RESEARCH REPORT SUBMITTED IN PARTIAL
FULFILLMENT OF THE REQUIREMENT FOR THE
DEGREE OF MASTER OF ENGINEERING**

**FACULTY OF ENGINEERING
UNIVERSITY OF MALAYA
KUALA LUMPUR**

2014

Abstract

Speaker identification is the mechanism of determining a person among a set of speakers to certify whether that person is who he claims to be. The available speaker identification systems are mostly based on the acoustical signal itself. The problem is that they are very sensitive to noise and can work only at very high signal-to-noise ratio (SNR). However, neural responses are very robust against background noise. In this study, a well-known model of the auditory periphery by Zilany and colleagues (J. Acous. Soc. Am., 2009) is employed to simulate the neural responses, known as neurogram, on identifying a speaker, and then average discharge rate or envelope (ENV) and the temporal fine structure (TFS) are computed from the neurogram. The resulted vectors are used to train the system by employing two types of classifiers, Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM). The database consists of text-dependent speech samples from 39 speakers, and 10 speech samples were recorded for each speaker in a quiet room. The performance of the proposed method is compared with the traditional acoustic feature (mel-frequency-cepstral-coefficient, MFCC) based speaker identification method for both under quiet and noisy conditions. As the neural responses are robust to noise, the proposed neural response based system using TFS responses performs better than MFCC-based method, especially under noisy conditions. In general, GMM shows better accuracy for the proposed method than using HMM as a classifier.

Abstrak

Pengenalan speaker adalah mekanisme menentukan seorang daripada satu set pembesar suara untuk mengesahkan sama ada orang itu yang dia mendakwa. Sistem pengenalan penceramah didapati kebanyakannya berasaskan isyarat akustik sendiri. Masalahnya ialah bahawa mereka adalah sangat sensitif kepada bunyi dan boleh bekerja hanya pada sangat tinggi isyarat-kepada-hingar (SNR). Walau bagaimanapun, tindak balas saraf yang sangat teguh terhadap hingar latar belakang. Dalam kajian ini, model terkenal pinggir auditori oleh Zilany dan rakan-rakan (J. Acous. Soc. Am., 2009) diambil kerja untuk mensimulasikan jawapan neural, dikenali sebagai neurogram, mengenal pasti penceramah dan kemudian kadar pelepasan purata atau sampul surat (ENV) dan struktur denda duniawi (TFS) dikira dari neurogram itu. Vektor menyebabkan digunakan untuk melatih sistem dengan menggunakan dua jenis penjodoh bilangan, Campuran Model Gaussian (GMM) dan Model Markov Tersembunyi (HMM). Pangkalan data ini terdiri daripada sampel ucapan teks yang bergantung daripada 39 pembesar suara, dan 10 sampel ucapan telah dicatatkan bagi setiap pembesar suara dalam bilik yang tenang. Prestasi kaedah yang dicadangkan dibandingkan dengan ciri tradisional akustik (mel-frekuensi cepstral-pekali, MFCC) penceramah kaedah pengenalan berasaskan kedua-dua di bawah keadaan tenang dan bising. Sebagai tindak balas saraf yang mantap dengan bunyi bising, sistem tindak balas berasaskan neural yang dicadangkan menggunakan balas TFS melakukan lebih baik daripada kaedah berasaskan MFCC, terutamanya dalam keadaan bising. Secara umum, GMM menunjukkan ketepatan yang lebih baik untuk kaedah yang dicadangkan daripada menggunakan HMM sebagai pengelas

Acknowledgement

I would like to thank Dr Muhammad Shamsul Arefeen Zilany, for giving me an opportunity to do this project under his supervision. It is also for his thorough guidance, brilliant ideas, and wide knowledge in the field which he passed along that enabled me to complete the task in doing the project in the given time.

My heartfelt gratitude goes towards Dr Wissam A. Jassim who patiently helped me to solve the problems arise in completing the project.

I would like to thank my adorable parents for their endless love and support, my beloved husband for his encouragement and devotion especially during doing this project, and finally my dear brother and sister.

Table of Contents

Abstrak.....	ii
Acknowledgement	iii
List of Figures and Tables.....	vi
Chapter 1. INTRODUCTION	1
1.1. Problem statement	2
1.2. Study Significance	3
1.3. Speaker identification	3
1.4. Human peripheral auditory system	4
1.5. Objectives	6
1.6. Scope of study.....	7
1.7. Outline of the report	7
Chapter 2. LITERATURE REVIEW.....	9
2.1. Auditory Nerve (AN) model	9
2.2. Model Description.....	14
2.3. Envelope (ENV) and Temporal Fine Structure (TFS)	16
2.4. Speaker Identification	16
2.5. Classification technique in speaker identification	19
2.5.1 Gaussian Mixture Model.....	20
2.5.2 Hidden Markov Model (HMM).....	26
Chapter 3. METHODOLOGY	30
3.1. Database.....	30
3.2. Preprocessing	31
3.3. Construction of Neurograms: AN model simulation.....	32
3.4. Overall design of system	32
3.5. Training using classification technique	34
3.6. Testing using probability density function (PDF).....	36
3.7. Calculation of system accuracy.....	38
Chapter 4. RESULTS and DISCUSSIONS.....	39
4.1. Results using GMM as a classifier	39
4.2. Speaker identification results using HMM as a classifier	40

4.3. Comparison of the performance of the proposed system with a MFCC-based speaker identification system	41
4.3. Effect of parameters	44
Chapter 5. CONCLUSION.....	46
5.1. Limitations	47
5.2. Future study.....	47
REFERENCES.....	49

List of Figures and Tables

Figure 1.1: Human auditory system	4
Figure 1.2 : Pictorial representation of basilar membrane.....	6
Figure 2.1: Model for significant nonlinear behaviors in the presence of noise	10
Figure 2.2: Block diagram of the Non-linear tuning properties model	11
Figure 2.3: OHC and IHC block model block diagram.....	12
Figure 2.4: C2-C1-Synapse Model	14
Figure 2.5: One dimensional probability density function pdf	20
Figure 2.6: A mixture of Gaussians distribution.....	21
Figure 2.7: One dimensional combination of mixture Gaussian distribution and pdf	21
Figure 2.8: Two dimensional Gaussian mixture distributions.....	22
Figure 2.9: Contour line for the mixture of the function.....	23
Figure 2. 10: Full covariance	24
Figure 2. 11: Diagonal covariance	25
Figure 2.12: a) correct number of K b) incorrect number of K	25
Figure 2.13: A Markov chain with three states	27
Figure 2.14: HMM state and transition random variables	28
Figure 3.1: Block diagram of speaker identification system.....	30
Figure 3. 2 a: Speech signal before pre-processing b: Speech signal after pre-processing	31
Figure 3.3: Flow chart of speaker identification system.....	34
Figure 3. 4: Flow chart of training phase of speaker identification system.....	36
Figure 3. 5: Flow chart of testing phase of speaker identification system	37
Figure 4.1 : Performance of the proposed system using ENV and TFS along with GMM as a classifier	40
Figure 4.2: Performance of the proposed system using HMM as a classifier.....	41
Figure 4.3: system accuracy comparison for GMM and HMM with the MFCC	42
Figure 4.4: Effect of the number of GMM components on the accuracy of the ENV-based speaker identification system	44
Figure 4.5: effect of changing GMM components on system accuracy for different levels of noise	45
Table 2.1: The probabilities that you carry the umbrella in different weathers.....	28

Chapter 1. INTRODUCTION

The speech signal contains different kind of information. The most important part is the message it conveys and then the identity of the speaker. The speech recognition deals with understanding the verbal message utterance, and in speaker recognition obtaining the identity of the speaker is the main responsibility of the system. In both cases, the speech can be specific words (text-dependent) or it can be different words (text-independent).

Speaker identification process consists of two main stages. In the first stage, a specific identity is assigned to a speech sample, and then in the second stage, authentication is done for a claimed speaker. The functionality of common speaker recognition and identification systems is mainly based on acoustical analysis of sound signals. However, the performance of these methods degrades substantially when background noise is added to the speech signal, meaning that they are not robust to noise. On the other hand, human performance of speaker identification far exceeds any system available today for both under quiet and noisy conditions.

Human sensory systems like eyes, ears and skin are in charge of communication to the outside world; they constantly receive stimulus and convert them to neurological impulses which can be interpreted by brain. The auditory part, as one of the important sensory organs of the body, keeps receiving sound pressure waves and converts them to chemical and electrical signals (action potentials). Auditory nerve is a complex sensory system that receives information acoustically, and it plays an essential role in the process of learning in our daily lives.

For understanding the procedure of signal transformation through the auditory pathway, it is strongly required to know about physiological process of auditory pathway. Physiological process of human auditory system is very sophisticated and has great abilities like understanding the direction of sound source or distinguishing original signal from background noise. This can be even extended to distinguish several speakers when they are talking at the same time. In another word, it acts very intelligently and is very robust to noise. This fact has given motivation to many researchers to work on implementing the models that can imitate the strategy employed by the auditory system for the same task.

1.1. Problem statement

Nowadays with rapid spreading of digital technology in almost every aspect of human lives, remote authentication as a feature has become of great need and applications, like security enforcement for banking and E-commerce, etc. But the inevitable property of all these applications is the increase of susceptibility to transmission channel noise. The performance degrades substantially under noisy conditions when the system is trained under quiet (clean) condition whereas tested with background noise. Current speaker identification systems are mainly based on the step by step analysis of the acoustical signal itself which are very sensitive to noise. So improving and making these systems robust to noise would be of great importance and need. Since human performance is very robust under diverse conditions, employing the same mechanism using a physiologically-based model of the auditory system might produce similar results.

1.2. *Study Significance*

Human auditory system has a great ability to distinguish and identify a speaker under noisy conditions. Based on this, researchers have tried to propose models of the auditory system to simulate the responses of the auditory nerves or neurons. On the other hand, the need for remote authentication systems like security enforcement and internet banking is growing rapidly. Another important application is the design and implementation of hearing-aid systems that can significantly improve the lives of people with hearing loss. So, the motivation of this project comes from this idea that employing an accurate model of the auditory periphery would improve the performance and accuracy of the speaker identification process, especially under noisy conditions.

1.3. *Speaker identification*

Speaker recognition is a biometric modality that uses a person's voice for identification purpose. Since speech contains much information that is unique for each person, the system should be able to recognize who is speaking without the need of getting information about the speaker at the same time. Speaker recognition is not same with speech recognition because the latter is about recognizing the words while they are pronounced, not the speaker.

Speaker recognition is further subdivided into two fields, speaker verification (SV) and speaker identification (SID). Speaker verification is certifying whether the person is whoever he/she claims to be and compared to the one stored in the system. But speaker identification, which is the topic of this research, is about determining an unknown speaker's identity by comparing his speech signal to a list in the database.

A speaker identification system consists of two parts. The first part is enrollment which is when the voice recording is done in order to form a model, some features are extracted. Then in the verification or testing process a speech sample or utterance is compared to all available models in the database. Then the best matched sample is determined to identify or verify the speaker.

1.4. Human peripheral auditory system

The auditory system comprised of three main parts, outer ear, middle ear and inner ear. Outer ear consists of earlobe (pinna) and auditory canal. Once the sound wave hits the pinna, it is reflected and attenuated, so this provides additional information for brain to know about the direction of the sound source. Sound wave then goes through auditory canal, and those waves which are between 3 and 12 kHz are amplified. Figure 1.1 illustrates the human auditory system.

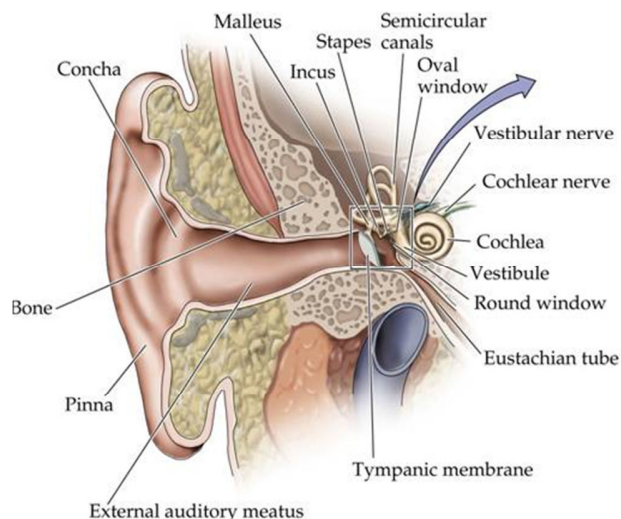


Figure 1.1: Human auditory system (Retrieved from <http://www.rci.rutgers.edu/>)

Middle ear consists of eardrum, three tiny bones (Ossicles) and oval window. The sound wave hits eardrum (tympanic membrane) located at the end of auditory canal and then travels across the air-filled middle ear cavity and Ossicles. The Ossicles are named as Malleus, Incus and Stapes. The main job of the Ossicles is to couple sound energy from the eardrum to the oval window by converting the lower-pressure eardrum sound vibrations into higher-pressure sound vibrations at the oval window. The reason of higher pressure is that the inner ear beyond oval window contains liquid rather than air.

Inner ear consists of the cochlea. The high pressure sound vibrations which are coming from the oval window move the fluid inside the cochlea and bend the neural receptors, such as Outer Hair Cells (OHC) and Inner Hair Cells (IHC). These neural receptors which are connected to the basilar membrane generate nerve signals and send them to the brain.

What basilar membrane does in auditory system is very similar to a filter. Different segments along the basilar membrane respond preferentially to different range of frequencies depending on the position of that particular segment. In other words, the auditory nerve shows a maximum response to a particular frequency, referred to as Characteristic Frequency (CF) based on the corresponding position on the basilar membrane. Higher frequencies are represented by the nerves attached on the base of the cochlea whereas lower frequencies are represented by the Apex. This can be further understood by referring to the Figure 1.2. In this study, a model for auditory responses has been used to mimic the processing strategy employed by the human auditory system to identify speakers under diverse background conditions.

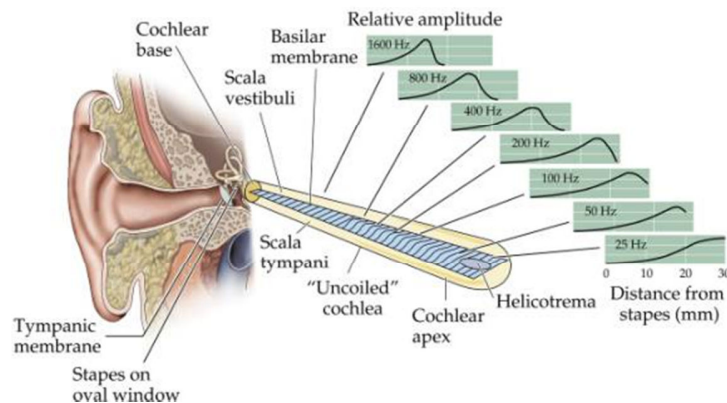


Figure 1.2 : Pictorial representation of basilar membrane (Retrieved from <http://www.rci.rutgers.edu/>)

1.5. Objectives

The goal of this project is to develop a neural response-based speaker identification system such that its performance under noisy condition would be better than traditional systems. Most of the traditional available systems today, usually use the features from the acoustic signal itself and are thus very prone to noise. However, in this study, an accurate model of the auditory periphery proposed by Zilany and colleagues (2009) will be employed. This model has proven to be capable of simulating the nonlinear behaviors of the AN responses. So it is expected that results of this project would be comparable with the performance of human subjects.

The objectives of the project are to:

- Develop a neural response based speaker identification system.
- Evaluate and compare the performance of different classification techniques.
- Test the performance of the system under noisy conditions.

- Compare the proposed system's performance with a traditional method such as Mel Frequency Cepstral Coefficient (MFCC) based speaker identification.
- Identify the best set of parameters in the classification techniques can be used.

1.6. Scope of study

In this study, a text-dependent speaker identification system is developed which is based on neural response. The database consists of thirty nine speakers and there are also ten speech samples available for each speaker. The speech samples are preprocessed by removing silence periods between utterances. Then the AN responses to these samples are simulated using the AN model proposed by Zilany and colleagues (2009). Two most widely used classification techniques, GMM and HMM, are used in this project to train the system. Finally the system is tested by using the GMM/HMM model for each speaker. For assessing the accuracy and robustness of the system, different levels of white Gaussian noise are added to speech samples, and Signal to Noise Ratio (SNR) is varied from -5 dB to +20 dB in steps of 5 dB. The performance of the proposed system is compared to the MFCC-based speaker identification system using the same classification technique. The effect of parameters for the classification technique is also evaluated.

1.7. Outline of the report

This report has five chapters. An introduction about speaker identification systems and human peripheral auditory system is provided in chapter one. The objectives and scope of this study is also included in this chapter. A literature review on AN model and its background is discussed in details in chapter 2. The basic theory of GMM and HMM as classification techniques is also covered. Chapter three is about the methodology of the

proposed speaker identification system. Different phases of the project are elaborated step by step, and the process of each phase is discussed in details. All the results are presented in chapter four with required explanations and discussions. Finally, chapter five provides the conclusion of the findings from this project, along with the discussion about limitation and potentials for future work.

Chapter 2. LITERATURE REVIEW

This chapter will discuss on previous studies on developing the AN model and speaker identification system based on AN model and classic methods. The theory of two classifications techniques, GMM and HMM is also presented in this chapter.

2.1. Auditory Nerve (AN) model

In order to understand the process done in auditory periphery system, using a model of auditory nerve (AN) fiber response can be an effective tool. So far many researchers have tried to establish a computational model. Flanagan and colleagues in 1960 tried to establish a model which can estimate the basilar membrane displacement by knowing the sound pressure at the eardrum (Flanagan, 1960). But their initial hypothesis was wrong because they had considered the cochlea as linear and passive. The initial models took account the properties of cochlea responses using only simple stimulus like clicks, single tones, or pair of tones (Geisler, 1976; Hall, 1981; Pfeiffer & Kim, 1973). Then some researches were done on modeling the cochlear responses to more complex sounds like speech. These models are relatively simple but none of them modeled the nonlinearities of basilar membrane (Lyon, 1984)

Later in 1987, Deng and Geisler reported significant nonlinearities in the responses of auditory nerve fibers to the speech sound. They described the main property of discovered nonlinearities as “synchrony capture” which means that the response produced by one formant in the speech syllable, is more synchronous to itself than what linear methods predicted from the fiber’s threshold frequency tuning curve (FTC). Their attempt was to explain such distinguished nonlinearities and also considering them in developing

new model. Their effort led to proposing a composite model which incorporated either a linear basilar membrane stage or a nonlinear one (Deng & Geisler, 1987).

Employing AN model as a front-end in many applications, such as modeling the neural circuits in the auditory brain stem (Hewitt & Meddis, 1993) or design of hearing-aid amplification schemes (Wilson et al., 2005) attracted the attention of some researchers. Patterson and his colleagues (1995) used different approach for modeling the AN. They paid attention on model's output itself in the form of auditory imagery. In their model, the processing of produced sound is done in form of graphical image of itself (Patterson, Allerhand, & Giguere, 1995). In 1999, Robert and Eriksson proposed a model which could imitate many events seen in the electrophysiological recordings related to tones, two-tones, and tone-noise combinations. Their model was able to generate significant nonlinear behaviors such as compression, suppression, and the shift in rate-intensity functions when noise is added to the signal (Robert & Eriksson, 1999). The Figure 2.1 shows this model.

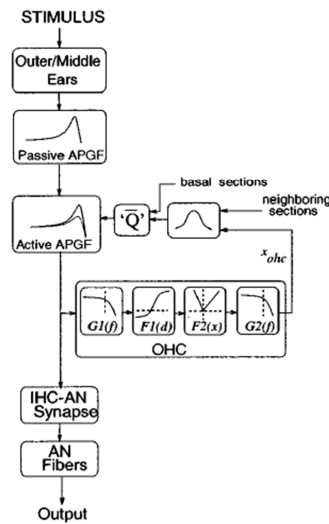


Figure 2.1: Model for significant nonlinear behaviors in the presence of noise (Robert & Eriksson, 1999).

Only two years later Zhang et al. (2001) presented a model which had some general features same as the model presented by (Robert & Eriksson, 1999). It consists of different parts, each providing a phenomenological explanation of a particular section of the cochlea function. They tried to address some important response properties that was not accomplished in the previous study such as temporal response properties of AN fibers and the asymmetry in suppression growth above and below CF. The new model focused more on nonlinear tuning properties like the compressive changes in gain and bandwidth as a function of stimulus level, the associated changes in the phase of phase-locked responses, and two-tone suppression (Zhang, Heinz, Bruce, & Carney, 2001). The schematic diagram of the proposed model by Zhang et al. (2001) is shown in Figure 2.2.

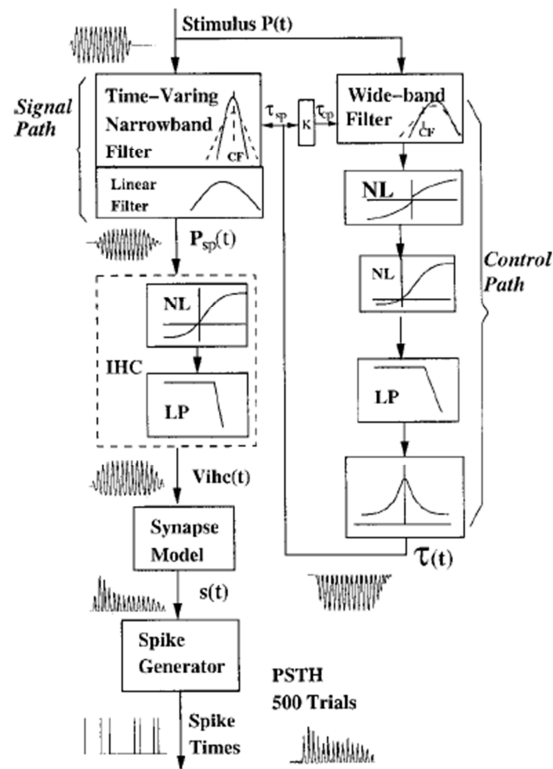


Figure 2.2: Block diagram of the Non-linear tuning properties model (Zhang et al., 2001).

Bruce et al. (2003) expanded the foresaid model of the auditory periphery to assess the effects of acoustic trauma on AN responses. They did some changes to model OHC and IHC and also made some modification to increase the accuracy in predicting responses to speech sounds. The modifications related to OHC led to model the effects of acoustic trauma, like various degrees of elevation and broadening of tuning curves and a proportional loss of compression and suppression. A similar modification to the IHC led to elevation of the tuning curve without substantially changing the bandwidth as measured by Q10 values. They concluded that both IHC and OHC impairment can cause model fibers with BFs near the second and third formants to lose synchrony to those formants and become more synchronized to other components of the vowel spectrum, same as what observed in the physiological data (Miller et al., 1997; Wong et al., 1998). However, their study was limited to low and moderate level responses (Bruce, Sachs, & Young, 2003). The schematic diagram of their model is shown in Figure 2.3.

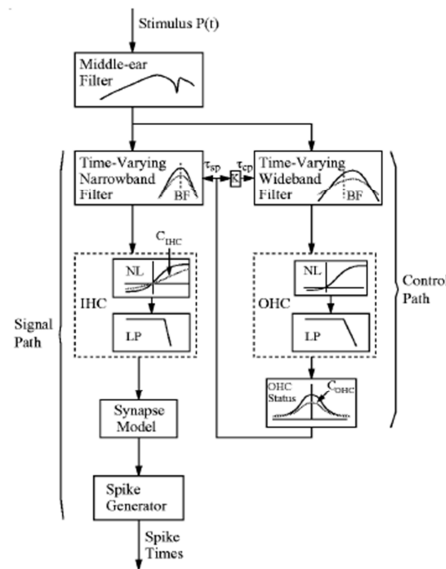


Figure 2.3: OHC and IHC block model block diagram (Bruce et al., 2003).

Zilany and Bruce (2006) presented a model to simulate normal and impaired AN fiber responses in cats. They extended the model presented by (Bruce et al., 2003) to also account for high level AN responses. This was accomplished by suggesting that inner hair cell should be subjected to two modes of basilar membrane excitation, instead of only one mode. Two parallel filters named C1 (component 1) and C2 (component 2) generate these two modes. Each of these two excitation modes has their own transduction function in a way that the C1/C2 interaction occurs within the inner hair cell. The transduction function was chosen in such a way that at low and moderate sound pressure levels (SPLs), C1 filter output dominated the overall response from the IHC output, whereas the high level responses were dominated by C2 responses. This property of Zilany-Bruce model makes it more effective on wider dynamic range of SPLs compared to those of previous AN models (Zilany & Bruce, 2006).

Zilany and colleagues continued working on their last model, but this time their focus was on the synapse model between inner hair cells and auditory nerve fibers. They included both exponential and power-law dynamics in their new model of rate adaptation at the synapse. Some recent evidences showed that the dynamics of biological systems can be described better by power law dynamics over the long-term whereas the exponential adaptation can capture the dynamics over the short time courses (Zilany, Bruce, Nelson, & Carney, 2009). Since the AN model used in this study is the one proposed by (Zilany & Bruce, 2006), a detailed description of model is given in following section.

2.2. Model Description

Since the AN model used in this study is the one done by (Zilany & Bruce, 2006), a detailed description of model is given in this section:

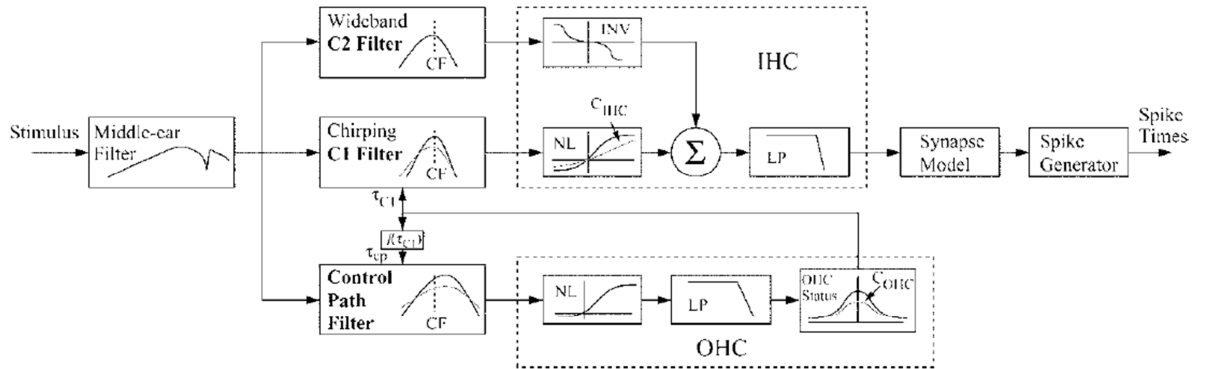


Figure 2.4: C2-C1-Synapse Model (Zilany & Bruce, 2006).

As it is illustrated in the Figure 2.4, the model consists of 7 blocks; middle ear, feed forward control path, C1 filter, C2 filter, IHC, OHC Synapse Model and Discharge generator. Each block's function corresponds to a specific component's function of the auditory-periphery system.

The input signal is an instantaneous pressure waveform in Pascal (Pa) which enters into the system through a middle-ear filter. The middle ear block is followed by two paths, signal path consisting of C1 and C2 filters and a feed forward control path. The job of control path is to adjust the gain and bandwidth of the C1 filter to account for different level dependent properties in the cochlea (Zilany & Bruce, 2006).

C1 filter: Is a narrow-band chirping filter. It replicates the tuning properties of the basilar membrane responses which are used as an input to the C1 IHC transduction function, and it accounts for low and moderate level responses.

C2 filter: C2 which is parallel to C1 is a static, linear and broadly tuned filter. The operation of C2 is based on Kiang's two-factor cancellation hypothesis (Kiang, 1990). It means that the transduction function followed by C2 will be affected by the level of stimuli and the output of transduction function is based on sound pressure levels. The hypothesis states that the interaction between the two paths produces effects such as the C1/C2 transition and peak splitting in the period histogram.

IHC: IHC transduces the basilar membrane response, which is in the form of mechanical displacement to electrical impulses or neural code. There are two types of IHC stereo cilia available in the peripheral auditory system, tall and short stereo cilia. C1 responses are produced by tallest stereo cilia and C2 responses are generated by shorter stereo cilia. In this model, simulation of IHC is done by adding C1 to the C2 responses which are 180° out of phase, and then the summed signal is passed through a low pass filter

OHC: The nonlinearities in the cochlea are introduced by the feed-forward control path. The function of the control path is to regulate the gain and bandwidth of C1 filter. However, according to the status of the OHC, a scaling factor, C_{ohc} , has been used in this path. Its range is between 0 and 1. The normal functioning is represented by $COHC=1$, and $COHC=0$ indicates to complete impairment.

Synapse model: This block simulates the IHC and auditory nerve fibers' synapse responses and its job is to determine the spontaneous rate, adaptation properties, and rate-level function of the AN model.

Spike generator: The input for this block is the synapse output which drives a non-homogeneous Poisson process and generates the spike timing.

2.3. Envelope (ENV) and Temporal Fine Structure (TFS)

As mentioned earlier, similar to spectrogram, Neurogram is a pictorial representation of a signal and it is used to visualize the output of AN model in the time and frequency domain. The colors correspond to the activity of the AN fibers in response to acoustic signal. Based on time resolution, two types of neurogram can be constructed such as envelope (ENV) and Temporal Fine Structure (TFS). The function of ENV or temporal envelope is to average the PSTH output of AN model intensity at each CF over a number of time frames and the speech is represented by a smoothed average discharge rates (Hines & Harte, 2012).

2.4. Speaker Identification

In the last few decades, several studies have been done based on using auditory model response for the purpose of speaker identification.

In 2011, Togneri and Pullella did a survey to improve robustness and accuracy of speaker identification system. According to them, feature vector extraction is that most basic process that can be applied by all kind of speaker identification. Linear Prediction Co-efficients (LPCs) are among the most common features which are directly extracted

from speech production model (Atal, 2005). Another common feature is PLP or Perceptual Linear Prediction coefficients (Mammone, Zhang, & Ramachandran, 1996).

Employing Fourier Transform also has become a popular approach for speaker identification task in the past few decades. This approach provides spectral based features like Mel-frequency spaced cepstral coefficients or MFCCs (Togneri & Pullella, 2011).

They also categorized the solutions for improving robustness to noise in three groups of feature-based, score-based or model-based. In the first group, the noise is discarded from the speaker information directly. In second group, the classifier scores are changed and in third group the distortion characteristics is combined with the speaker models. Cepstral mean normalization technique (Furui, 1981), RASTA processing (Hermansky, Kohn, Morgan, & Bayya, 1992) and warping methods (Pelecanos & Sridharan, 2001) are some examples of the studies related to feature-based group. The disadvantage of above technique is that they have to be matched to the environmental disturbances and parameters like the stationary of noise effect on it (Togneri & Pullella, 2011).

Shao and Wang (2007) got advantages of using auditory feature uncertainties to improve the robustness of speaker identification system. Based on auditory periphery model they introduce two new features. The first feature is Gammatone feature (GF) which is drawn out of a bank of filters called Gammatone filters. Gammatone filters were initially applied to simulate the human cochlea filtering process. The other new feature is Gammatone frequency cepstral coefficients or GFCC which are derived from the first feature, GF.

In this study, authors employed a method presented by them previously, called missing data method. Their purpose was to improve robustness of speaker identification and speaker verification task under noisy condition. The main idea of this technique is to break down the speech signal in time-frequency units and then the noisy part of speech signal in time-frequency is treated as a missing data. Afterwards a binary mask is needed to determine if a specific time-frequency unit is a missing unit or not. Finally they could reconstruct the missing GF by using a priori data which was derived from the speech training set (similar to using a Universal Background Model).

In order to obtain the GFCC features, they applied discrete cosine transform (DCT) to the Cepstral domain of GF. Same as the way MFCC is obtained in spectral analysis. Then GMM classifier is used to train the system. Eventually the accuracy of the proposed GFCC-based features is compared to other features which happened to be the highest at -6 dB SNR level (Shao, Srinivasan, & Wang, 2007).

In the study done by Abuku et al (2010), a feature vector was extracted and employed for the purpose of speaker identification. This feature vector is based on post-stimulus time histogram (PSTH) of AN model output signal. The AN model output signals are in the form of multi-dimensional pulse signals. This AN model is based on Meddis (1986) IHC model which is enhanced with phase-locking model, proposed by Makai et al (2009) (Meddis, 1986). In this study, the speaker identification process is performed using Japanese vowels where the database consists of 12 speakers. In order to increase the accuracy of system, standardization and normalization are done as additional steps by authors. They compared their results to conventional method based on by using LPC analysis. Applying standardization and normalization of the PSTH output makes the

average of the speaker identification accuracy becomes higher than the accuracy obtained from LPC analysis (Abuku, Azetsu, Uchino, & Suetake, 2010; Azetsu, Abuku, Suetake, & Uchin, 2012).

Another study about speaker identification was conducted by Li, Qin (2010). They tried to propose a new feature based on their own model of peripheral auditory hearing system, implemented in 2003. The new feature is expected to overcome the disadvantages of FFT and its inverse transform. They proposed a new technique called Auditory Transform (AT) which is an invertible auditory-based transform. This is done by using Gammatone filter as the cochlea model. AT contains both forward and inverse transform. In the forward transform, the decomposition of speech signal into a number of frequency bands is done while in the inverse transform the original signal is rebuild. The advantage of AT technique is the adjustability of filter bandwidth according to applications. The results shows better performance than MFCC based identification system (Li & Huang, 2010).

2.5. Classification technique in speaker identification

The job of classifiers in speaker identification process is to train the system from a set of observations. In other words, a classification technique acts as supervised machine learning process. In order to do that, a classifier requires several speeches of the speakers to train data during system set up.

A classifier is nothing but a mathematical algorithm based on statistical analysis. It works on raw data and gives outputs based on the highest probability of a feature vector to belong to a specified category. It actually maps a training set to particular group or clusters.

Depending on the type of the field in which classification is required, there are many kinds of classifiers for example Bayes classifier neural k-nearest neighbors networks, Support Vector Machine (SVM) Gaussian mixture model (GMM), Hidden Markov Layer(HMM) and etc. so far HMM and GMM technique are shown to be a successful approach in signal verification and identification filed.

2.5.1 Gaussian Mixture Model

Initially it is as a good idea to visualize GMM, to make it easier to understand. We know that a one dimensional random variable x , has a probability density function or PDF as it is illustrated in Figure 2.5. Now suppose there is more than one, or a mixture of Gaussians distribution, like what is shown is Figure 2.6.

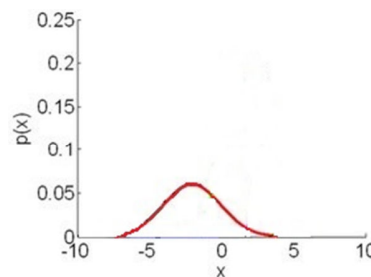


Figure 2.5: One dimensional probability density function pdf (Retrieved from <http://www.slideshare.net/>)

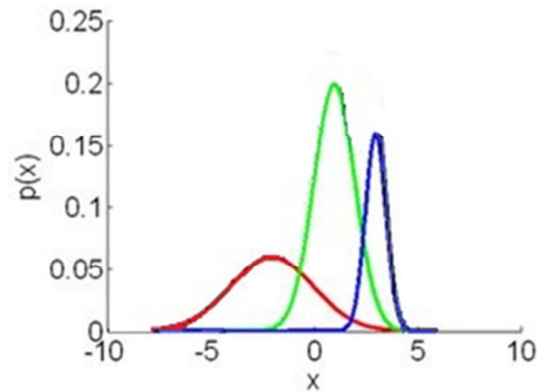


Figure 2.6: A mixture of Gaussians distribution (Retrieved from <http://www.slideshare.net/>)

In this case the probability density function for the mixture of Gaussian is going to be linear combination of these individual PDFs. The black line in Figure 2.7 shows this combination. Two dimensional Gaussian mixture distributions is also shown in Figure 2.8.

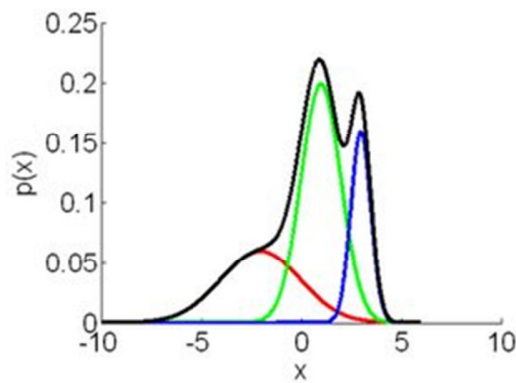


Figure 2.7: One dimensional combination of mixture Gaussian distribution and pdf (Retrieved from <http://www.slideshare.net/>)

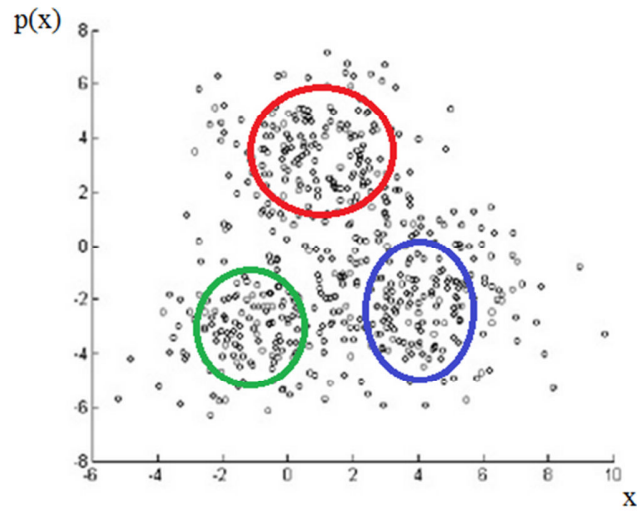


Figure 2.8: Two dimensional Gaussian mixture distributions (Retrieved from <http://www.slideshare.net/>)

The contour line for the mixture would be a single function just like the plot shown in Figure 2.9. So as is clear from the figures above, three different feature vectors observations of x (blue, green and red colors) are normally distributed based on their probability $p(x)$ and the overall probabilities can be represented by combining all three Gaussians into a single mixture of Gaussian density (black line) through its probability density function (PDF) of the original observation.

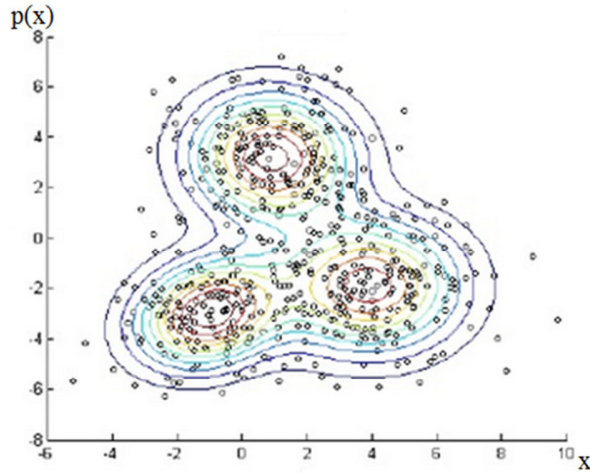


Figure 2.9: Contour line for the mixture of the function (Retrieved from <http://www.slideshare.net/>)

Back to the speech signals, each particular speaker has got several feature vectors. So similar to examples above, GMM is used to form a distribution for these feature vectors of each speaker.

Now going deeper into the theory, the mean, variance and deviation about the mean are three parameters by which the Gaussian densities are characterized. But when there is no such information, GMM algorithm has got no way to know which features belong to which Gaussian distribution. So in this case the combination of EM method should be added as an optimization of the Gaussian mixture that will maximize the likelihood of the observed data. GMM is the weighted sum of a number of Gaussians. The weight is determined by distribution as shown in Eq 2.1 and Eq 2.2.

$$p(x) = \pi_0 N(x|\mu_0, \Sigma_0) + \pi_1 N(x|\mu_1, \Sigma_1) + \cdots + \pi_k N(x|\mu_k, \Sigma_k) \quad \text{Eq 2.1}$$

Where:

$$\sum_{i=0}^k \pi_i = 1 \quad \text{Eq 2.2}$$

So it can be written in Eq 2.3:

$$p(x) = \sum_{i=0}^k \pi_i N(x|\mu_i, \Sigma_i) \quad \text{Eq 2.3}$$

Using GMM, provides a powerful tool for classification purposes, but like any other method, it has got its own disadvantages and limitations based on the type of application it is been used.

One of limitations is that in order to properly estimate the model's parameter, GMM requires enough data. This problem can be, to some extend and not completely, overcome by using different shapes of covariance matrix like full, diagonal, spherical etc. Figure 2.10 shows full covariance which is also the default setting of GMM in Matlab.

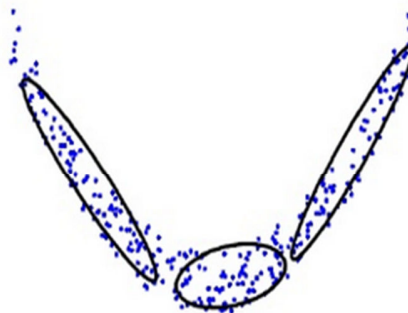


Figure 2. 10: Full covariance (Retrieved from <http://www.slideshare.net/>)

It can be seen that this shape fits data best, but in high dimensional space, it would be costly. To reduce the cost, the setting for shapes of covariance matrix can be changed to diagonal, shown in Figure 2.11.

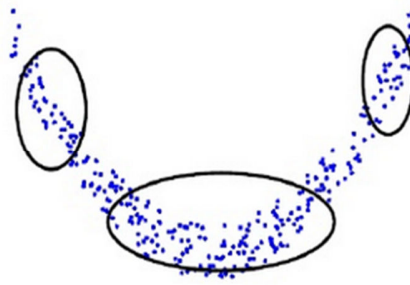


Figure 2. 11: Diagonal covariance (Retrieved from <http://www.slideshare.net/>)

In this case GMM can place the feature observations into a less specific area. But then there would be a need to tradeoff between the cost and quality. Because the quality will decrease if all observations are not covered (as in this case) (Reynolds, Quatieri, & Dunn, 2000; Togneri & Pullella, 2011)

Another limitation is regarding the number of Gaussian component (K). In order to get global likelihood for all training data that would not be specific, proper training is a must. So the best number for Gaussian component should be chosen so that the maximum likelihood on each iterations would be obtained (Reynolds et al., 2000; Reynolds & Rose, 1995). Figure 2.12 shows how selection of K effects on clustering the data.

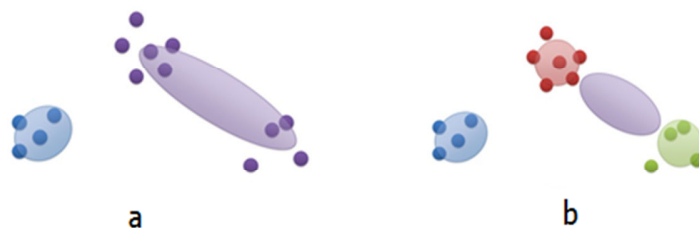


Figure 2.12: a) correct number of K b) incorrect number of K (Retrieved from <http://www.slideshare.net/>)

Finally the last disadvantage of GMM is when singularity happens. Sometimes in training phase some data are not seen but then in testing phase they appear so subsequently the performance of system is degraded. In speech signal processing field, one solution could be using text-independent speeches or using the words which contain all possible phonemes of the language. But this is not an ultimate solution because sometimes, just like current study, the goal of system is basically to work with text-dependents speeches. Including different training data and increasing their number is another solution but this also leads to lower the performance speed (Reynolds et al., 2000; Reynolds & Rose, 1995).

2.5.2 Hidden Markov Model (HMM)

Hidden Markov Model (HMM) is a generalized Bayesian statistical model which deals with a sequence of parameters. One of the most used application of HMM is speech recognition (Baker, 1975). HMM generally lets you predict the probability of an event in which there was not seen or observed. As an application, HMM has been widely used in speech recognitions. In this example the voice signal is the observed data however the speaker or the words and sentences are not observed or in another state they are hidden.

2.5.2.1 Discrete Markov process

By assuming a system in Figure 2.13, the system can be in any individual state such as S_1, S_2, \dots, S_N . The system's state is changing depend on a set of probabilities in discrete timing but constant. For instance in $t=1, 2, \dots$ the state in time (t) is shown by q_t . For an appropriate explanation of current state of the system the knowledge of all previous states is needed. In the Markov assumption it is important that the probability of current

state only depends on the previous state in the time. Eq 2.4 shows the last sentence in mathematics.

$$P(q_t = S_j | q_{t-1} = S_i, q_{t-2} = S_k, \dots) = P(q_t = S_j | q_{t-1} = S_i) \quad \text{Eq 2.4}$$

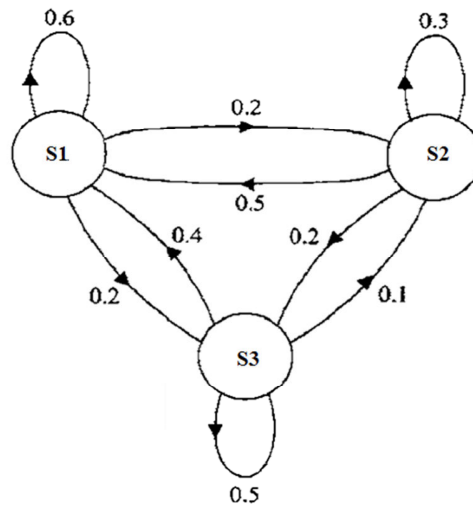


Figure 2.13: A Markov chain with three states (Retrieved from <http://www.powershow.com/>)

The general idea behind the HMM is shown in Figure 2.14. Each round shape is a random variable that can contain any values. $x(t)$ is a random variable in a hidden state and $y(t)$ is a random variable for observation. Both $x(t)$ and $y(t)$ are in the time (t) . Each arrow means the conditional dependencies between the variables. So as it is stated above, $x(t)$ is only depending on $x(t-1)$ and $x(t-2)$ has no effect on the x in time t . Similarly, $y(t)$ is only depending on the hidden variable $x(t)$, both in time t .

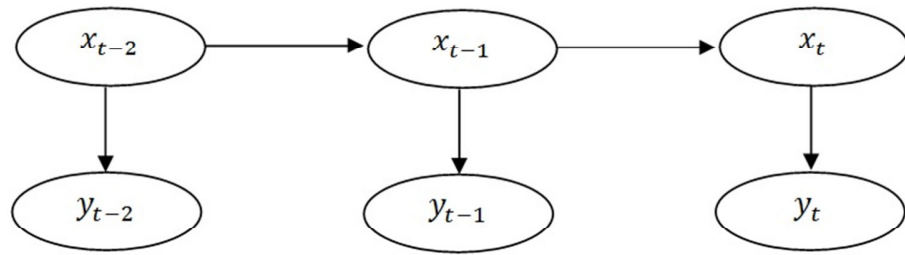


Figure 2.14: HMM state and transition random variables (Retrieved from <http://en.wikipedia.org/>)

There is a well-known example about the basic principal of HMM that assumes a robot in a house which is supposed to guess the weather outside according to a person's behavior. The only clue that it has, is whether that person is carrying an umbrella with him when he arrives home or not. It should be remembered that actual weather is hidden in this example. Now, the probabilities' assumption is as followed in Table 2.1.

Table 2.1: The probabilities that you carry the umbrella in different weathers

	Umbrella's probability
Weather is Sunny	0.2
Weather is Rainy	0.6
Weather is Windy	0.2

According to Markov process, the equation of weather changes probabilities is as Eq 2.5.

But since the weather is a hidden parameter for the robot now, based on Bayes' Rule in Eq 2.6:

$$P(w_1, w_2, \dots, w_n) = \prod_{i=1}^n P(w_i | w_{i-1}) \quad \text{Eq 2.5}$$

$$P(w_1, w_2, \dots, w_n | u_1, u_2, \dots, u_n) = \frac{P(u_1, u_2, \dots, u_n | w_1, w_2, \dots, w_n) P(w_1, w_2, \dots, w_n)}{P(u_1, u_2, \dots, u_n)} \quad \text{Eq 2.6}$$

Where u_i is a binary condition that he brings the umbrella with him or not. For example u_i is true if he brings and is false if he doesn't. In the Eq 2.6, $P(u_1, u_2, \dots, u_n)$ is the sequential observation's probabilities of seeing umbrella or not for example. And the denominator part of the Eq 2.6 also can be rewritten in $\prod_{i=1}^n P(u_i | w_i)$ format. By assuming that for all i , u_i and w_i are independent of u_j and w_j for all j , if $i \neq j$.

HMM has been implemented in so many different areas. HMM is useful when an algorithm is required to infer information from sequences that is not observed directly. Some of these applications are Speech recognition, Machine translation, Gene prediction, and Speaker detection, Bioinformatics, Predictions and Filtering.

Chapter 3. METHODOLOGY

This chapter describes the methodology of the neural response based speaker identification system and the detailed information of each step is given. Figure 3.1 illustrates the overall procedure in the form of block diagram. As it is shown, the process consists of four main stages which are discussed in detail through this chapter. The flow chart of the speaker recognition/identification system is also presented. The Matlab software has been used to simulate and analyze the responses in all parts of this project. (Matlab 2013a, The Mathworks Inc).

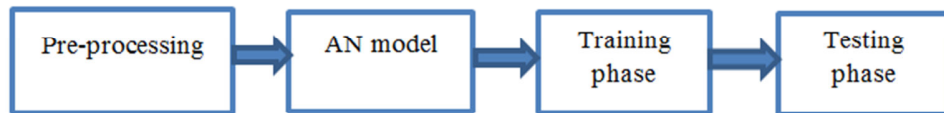


Figure 3.1: Block diagram of speaker identification system

3.1. Database

As mentioned earlier, the goal of this project is to identify a speaker among a set of known speakers in a database. The database used in this study consists of text-dependent speech samples from 39 speakers. They all were in the age of 22 to 24 years, and 25 of them were males and 14 were females. The recording was done in a quiet room using a microphone with 16-bit quantization rate, and a sampling rate of 8 kHz was used. The speakers were asked to say 'University Malaya' 10 times in different recording sessions, meaning that there are 10 different samples for each speaker. Seventy percent of the

available speech samples were used for training, and only thirty percent was used for testing.

3.2. Preprocessing

In this stage, the silence part of speech samples was removed by using VOICEBOX which is a speech processing toolbox in Matlab. The Figure 3. 2 shows the speech signal before and after preprocessing. The removal of silence period among the speech samples helped to align the utterances for each speaker and thus improve the results.

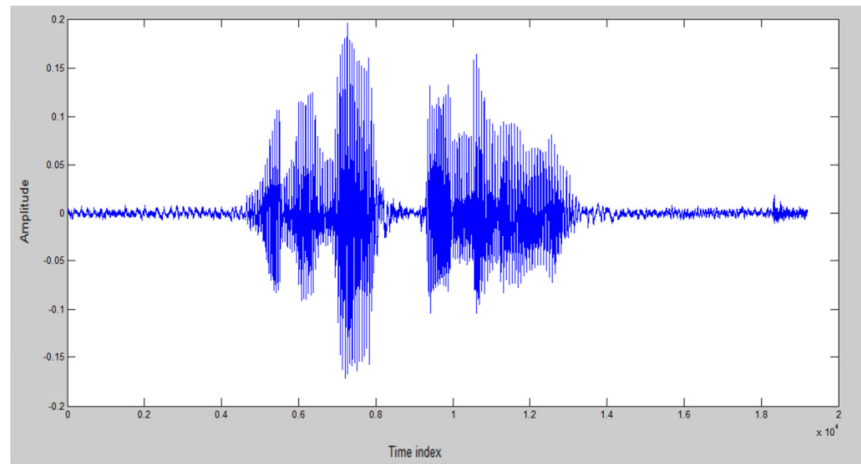


Figure 3. 2 a: Speech signal before pre-processing

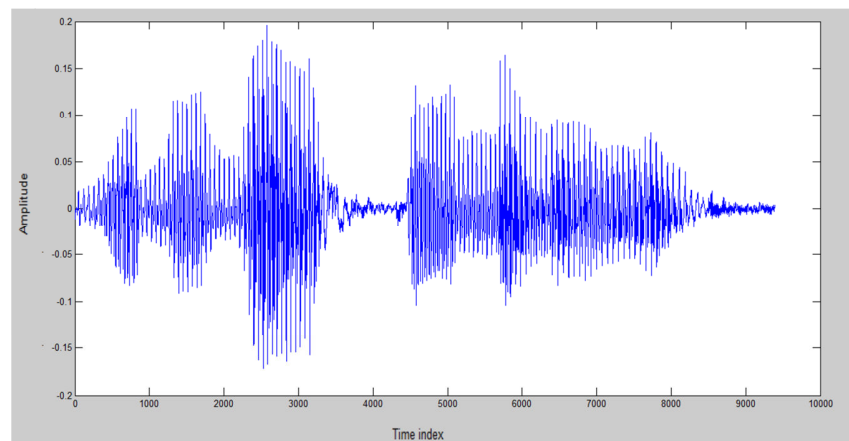


Figure 3. 2 b: Speech signal after pre-processing.

3.3. Construction of Neurograms: AN model simulation

The AN model used in this project is a well-established and known model proposed by Zilany et al. in 2009. This model has been used to simulate the neural responses for a range of Characteristic Frequencies (CFs) spanning the dynamic range of hearing. This model can successfully capture a wide range of nonlinear physiological phenomena, and the model has been validated against a wide range of physiological recordings from the literature. While programming, all the initial parameters of the AN model were set as default. All of the input signals were sampled at 100 kHz required by the AN model. Responses of the model were simulated for 32 CFs ranging from 250 Hz to 8 kHz; the CFS were logarithmically spaced over the entire range, mimicking the frequency spacing in the cochlea. Since the model responses were simulated for a normal listener, the scaling parameters C_{ohc} and C_{ihc} were set to 1.0. The model simulations were done for all 39 speakers, each with their 10 utterances. The results were saved for further processing.

3.4. Overall design of system

In order to understand the methodology of speaker identification, it is a good idea to pay attention to the difference between identification and verification process. In identification, test samples (here thirty percent for each speaker) were tested against the GMM models of all other speaker, while in speaker verification, the test samples are tested against the GMM model of only the claimed speaker. That's why speaker identification process is much slower than the speaker verification system.

The flow chart shown in Figure 3.3 can give a quick overview of how the speaker identification system works. In the first step, all speech samples are given as an input to the AN model, and subsequently model responses are simulated to produce ENV and TFS neurograms using synapse output (probability of spike rate as a function of time). Then 30% of the speech samples are chosen randomly for each speaker as testing samples. The rest 70% are used in training part to make the respective GMM model for each speaker. In the final step, the probability density function (PDF) is calculated for each testing sample using all GMM models. Then the highest value is found. The GMM model of corresponding to the highest value determines the identity of speaker. However, under noisy condition, it is possible that because of noise, the probability value of an irrelevant GMM model speaker becomes highest; in this case the identity of the speaker will be wrong.

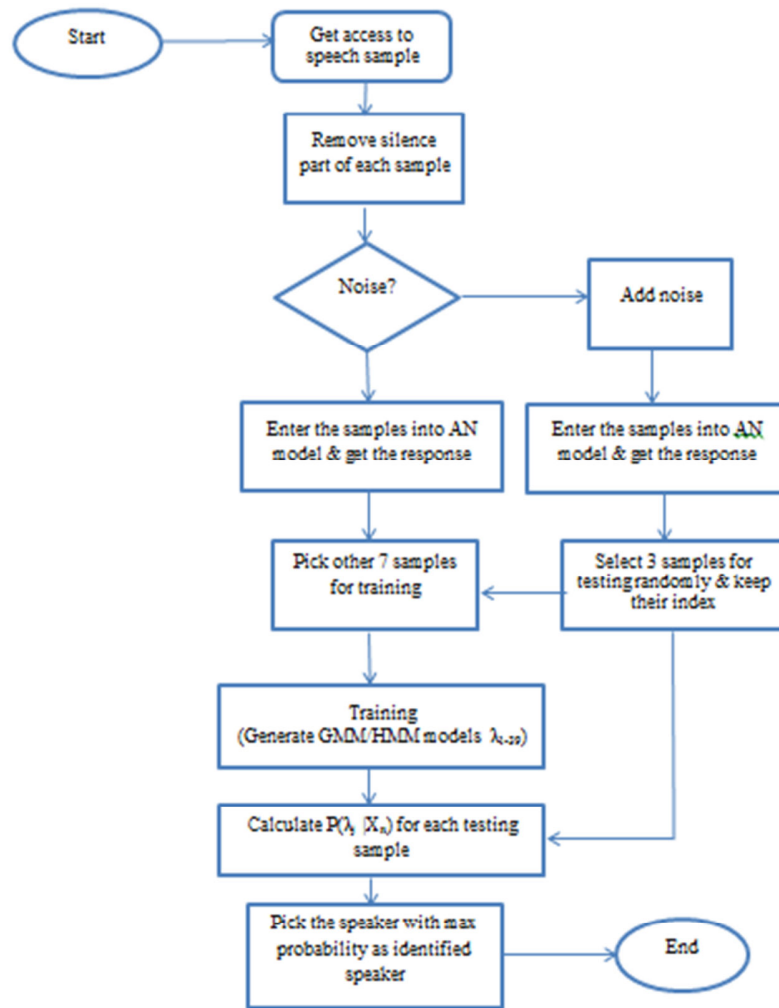


Figure 3.3: Flow chart of speaker identification system

3.5. Training using classification technique

There are many classification techniques available in speech signal processing. GMM and HMM are the common and oldest methods which are used in this project. Figure 3.4 shows the flow chart of the training part of the speaker identification system. Each speech sample is given as an input to the AN model, and the model generates the outputs correspondingly in .mat files and saves them in the Matlab environment. Initially,

the neurogram in response to each speech sample is a $(d \times n)$ matrix, where d corresponds to the number of characteristic frequencies and n is the number of data points in the speech sample. Based on time resolution, the synapse out is re-binned to construct ENV or TFS neurograms. The re-binning of $100\mu\text{s}$ followed by a 128 point smoothing using a Hamming window with 50% overlap results an ENV neurogram, and the maximum frequency content is ~ 160 Hz. On the other hand, re-binning to $10\mu\text{s}$ followed by a 32 point smoothing produces TFS neurogram, and the maximum frequency is ~ 6.7 kHz, which is the range of phase-locking to each individual cycles of the stimulus frequency.

As mentioned earlier, each speaker has 10 speech samples. Seven out of this 10 ENV/TFS samples are chosen to train the system. These seven samples are concatenated with each other to form a single matrix of $(N \times d)$, $[XN]$; with a fix number of $d=32$ and $N=n_1 + n_2 + n_3 + \dots + n_7$. For GMM, $[XN]$ is given as an input argument to “`gmdistribution.fit`” function in Matlab. And for HMM, $[XN]$ is given as an input argument to “`mhmm_em`” function in Matlab. The distribution value of components K for GMM is ranged as 4,8,16,20,25,32,64 and 128. This process is done for all speakers and the output of this function is saved in matrix `GMM_all_models/` `HMM_all_models` and later on these models will be used in testing stage.

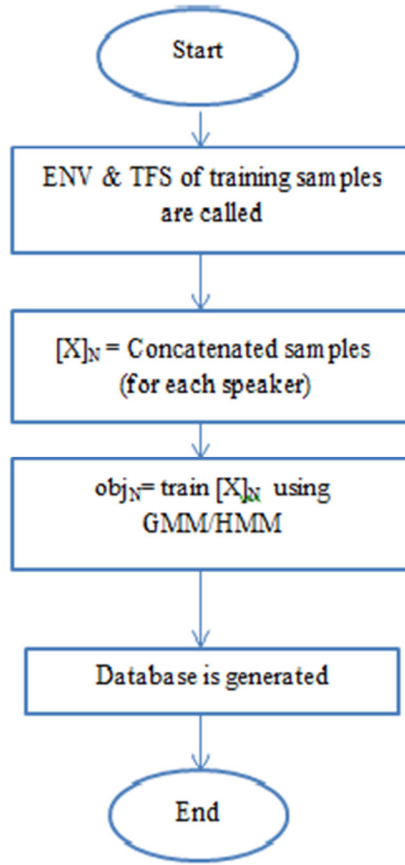


Figure 3. 4: Flow chart of training phase of speaker identification system

3.6. Testing using probability density function (PDF)

Figure 3.5 shows the flow chart of the testing stage for a speaker identification system. In this stage, the testing samples are applied to test the accuracy and validity of the system. This is done for all 39 speakers. In order to assess the system performance under noisy conditions as well as in quiet condition, the noise is added to all speech samples by the function “*awgn*” in the Matlab environment before doing AN model simulation. This function adds white Gaussian noise with different level of signal-to-noise ratio (SNR) ranging from -10dB to +20 dB in steps of 5 dB. The testing samples are chosen randomly.

The testing samples along with the GMM/HMM models (generated from training stage) are given as input to the PDF functions, and a vector is generated as output. For each testing sample of each speaker which is a $(n \times d)$ matrix, PDF function generates a vector of n by 1 , where each value of that vector represents the PDF for a given GMM model. Since PDF values are too small, logarithm of these values are calculated to make them in the range of minus infinity to one. In the next step, the mean of each vector is taken, and then the speaker model corresponding to the maximum value is selected as the identified speaker. The accuracy of the system is assessed in percentage. Figure 3.6 shows related block diagram as well.

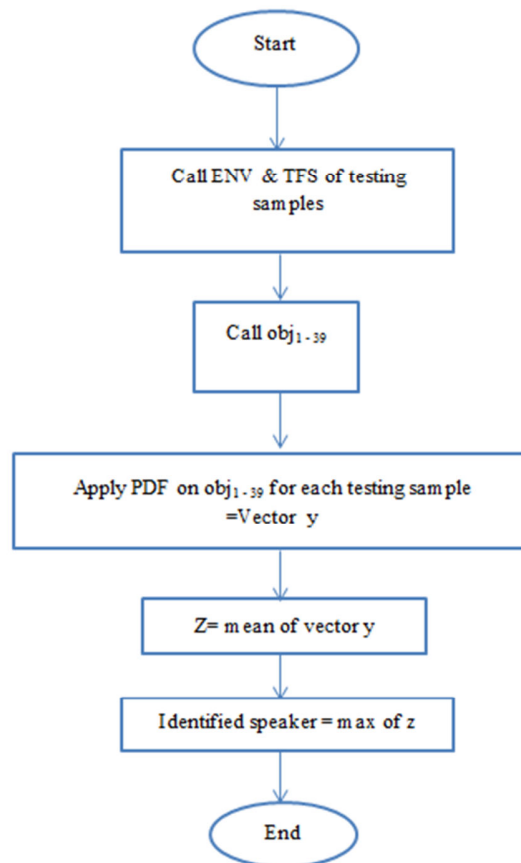


Figure 3. 5: Flow chart of testing phase of speaker identification system

3.7. Calculation of system accuracy

In the last phase of any process the performance is calculated. So similarly the accuracy of speaker identification system is calculated by making a matrix called confusion matrix. This matrix has 39 columns and 39 rows. The rows represents speakers starts from speaker number one to speaker number 39. Each column represents corresponding GMM/HMM model for each speaker means column one represents GMM/HMM model for speaker number one. So when the accuracy is 100% true, this matrix will be a diagonal matrix. At the end of program the accuracy is calculated as Eq 3.1.

$$\text{accuracy} = \frac{\text{sum of diagonal elements of confusion matrix}}{\text{total number of speakers} * \text{number of tests}} \times 100 \quad \text{Eq 3. 2}$$

In above equation total number of speakers is thirty nine and number of tests is three so the equation above will be:

$$\text{accuracy} = \frac{\text{sum of diagonal elements of confusion matrix}}{39 * 3} \times 100 \quad \text{Eq 3. 3}$$

Chapter 4. RESULTS and DISCUSSIONS

In this chapter, system performance using two different classification techniques, GMM and HMM is assessed and compared to the traditional method of MFCC. Since it is important to know which types of neurogram measurement contains more information about the identity of the speaker, both ENV and TSF neurograms were calculated. The system performance under noisy condition is evaluated by introducing different levels of noise to the original speech signal. There are several effective parameters in this study. Variation of each parameter in their appropriate ranges and also changing them with respect to one another can give an insight about the potential mechanism used by the auditory system.

4.1. Results using GMM as a classifier

In this study, the output of the AN model is subjected to a classifier. The performance of the proposed system is evaluated for two versions of the neurogram, ENV and TFS, and is shown in Figure 4.1. The results are shown as a function of SNR, which is varied from -5 to +20 dB in steps of 5 dB. The standard deviation of each point is also presented in the form of error bars. As far as GMM component number K is concerned, the performance of the system is plotted for their optimal value of K which is 20 for ENV and is 128 for TFS. In the following sections, it will be discussed how these optimal values for K have been obtained. The number of iteration is set to 500 times, and normalization is also done on the neurogram in order to improve the result. This condition is kept constant for all sections of this chapter.

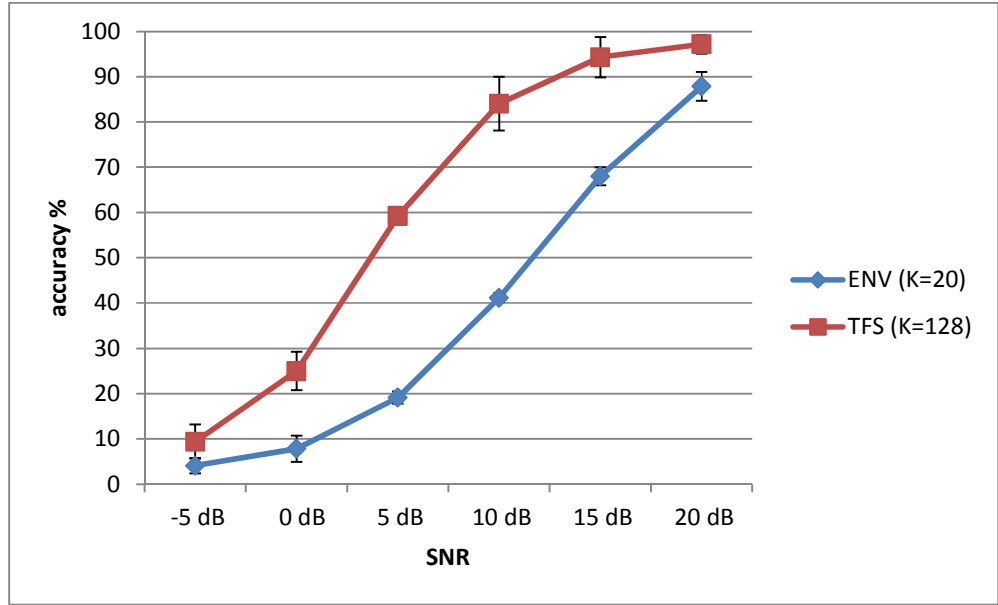


Figure 4.1 : Performance of the proposed system using ENV and TFS along with GMM as a classifier

In general, the performance of the proposed system declines as more and more noise is added to the speech signal, consistent with the results from the behavioral studies. Under quiet condition, speaker identification performance is nearly ~100%, whereas it drops to ~10% when a background noise of SNR -5 dB is used. Although the performance using TFS and ENV is comparable at very high and low SNRs, TFS information gives better speaker identification performance in the intermediate SNR levels (0 to 15 dB). This finding suggests that phase-locking information to the individual stimulus frequency is important for speaker identification, which is supported by the well-known fact that the difference in fundamental frequency plays a big role in speaker identification.

4.2. Speaker identification results using HMM as a classifier

In this section the performance of the proposed system is evaluated when HMM is used as classifier. As far as HMM parameters are concerned, the number of hidden states is set to 25 and the HMM is employed with mixture of Gaussians outputs. The neurogram

applied in this section is of ENV type. The system performance is plotted in Figure 4.2. It can be understood by figure that even in low noise level like 20 dB the performance of system is very poor. However this is a primary result for HMM classification technique.

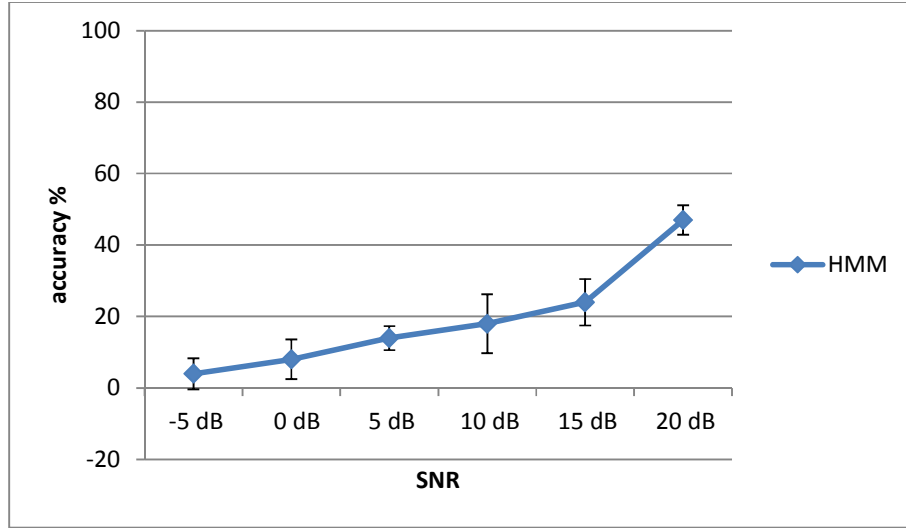


Figure 4.2: Performance of the proposed system using HMM as a classifier

4.3. Comparison of the performance of the proposed system with a MFCC-based speaker identification system

In this section, the performance of the proposed neural response-based method has been compared to a traditional acoustic feature (MFCC) based method. In both cases, GMM has been used as a classifier. The performance of the proposed system with HMM as a classifier is also shown. The performance has been evaluated for the same database (39 speakers with 10 samples from each speaker), and the number of GMM components, K , is 32 in this case. Regarding the MFCC based method; the VOICEBOX toolbox has been used to calculate MFCC. For derivation of MFCCs, at first the Fourier Transform of a windowed excerpt of signal is taken. Then by using triangular overlapping windows, the

powers of the spectrum obtained above is mapped onto the mel scale. In the next step the logs of the powers at each of the mel frequencies is taken. After that the discrete cosine transform of the list of mel log powers is taken, as if it were a signal. Finally the MFCCs are the amplitudes of the resulting spectrum.

Regarding the number of coefficients, there are 12 coefficients per each frame. The frame size is $0.03 \cdot f_s$, where f_s is the sampling frequency of 8 kHz. The frames are extracted using Hamming windows with 50% overlap.

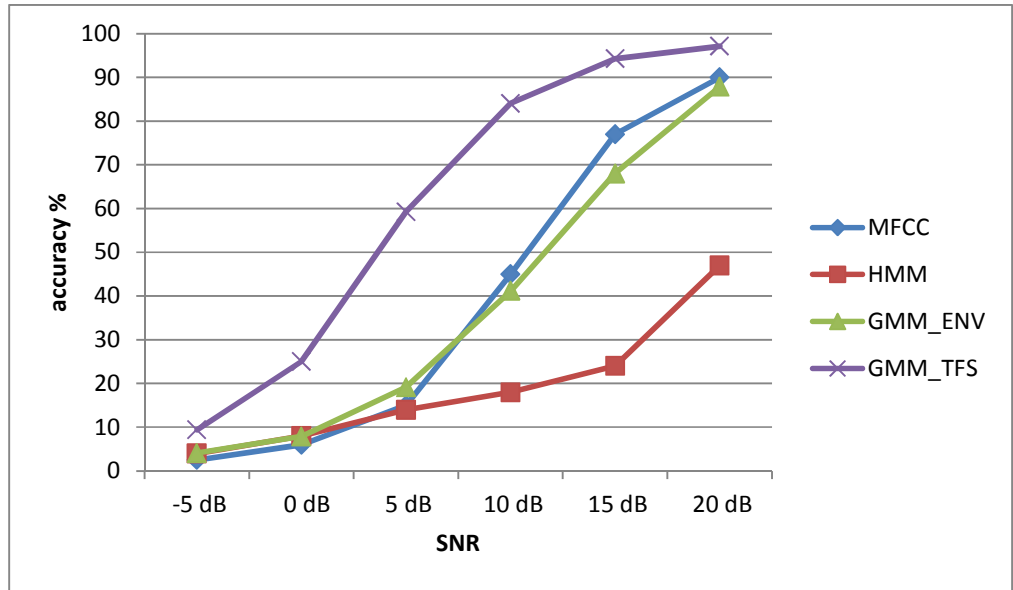


Figure 4.3: system accuracy comparison for GMM and HMM with the MFCC

Figure 4.3 shows the performance of the proposed and MFCC-based method as a function of SNR. Among all of the conditions, the performance of the neural response based method with TFS neurogram and GMM as a classifier is the best at all SNRs. Its

performance reaches to 98% at an SNR of +20 dB, and for SNR in the range of 0 to 10 dB, there is a noticeable difference between the performances of TFS based method and other methods. It is also noticeable that the performance of the ENV-based method is very comparable to the results from the MFCC using GMM as a classifier. Among the all methods, the performance of the neural response-based speaker identification using HMM as a classifier is poorer such that the performance is ~50% even under quiet condition. However, the performance is comparable to a MFCC- or ENV-based method (using GMM) for SNRs of 5 dB and below.

4.3. Effect of parameters

When GMM is used as a classifier, it is important to notice that the performance is substantially affected by the number of GMM component (K). In this section, the system accuracy is evaluated for a range of GMM components K , varying from 4 to 128. The effects have been evaluated for both ENV and TFS based methods using GMM as a classifier, and the results are shown in Fig. 4.4 and 4.5, respectively.

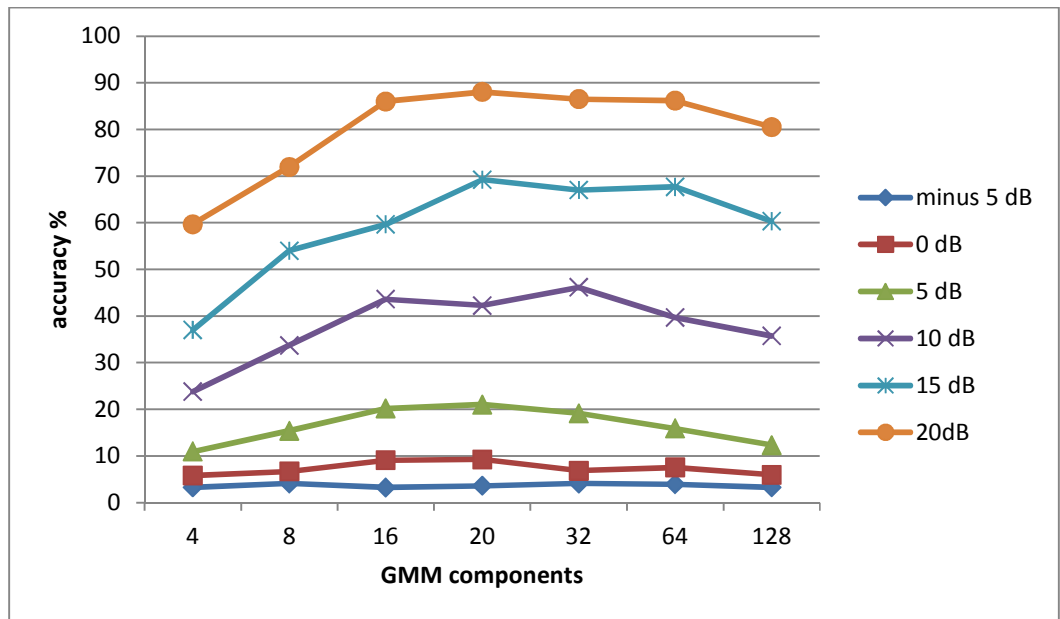


Figure 4.4: Effect of the number of GMM components on the accuracy of the ENV-based speaker identification system

As Figure 4.4 shows, the GMM component numbers ranging from 16 to 64 seem to provide the highest accuracy. Being more specific, it can be said that for most of SNR values (other than 10dB) the K value of 20 gives the best result. However, the overall fluctuation of accuracy value for all SNR level in the range of K equals to 16 to 64, is not noticeable (average = 4%).

Figure 4.5 shows the effects of number of GMM components (SNR as a parameter) on the performance of TFS-based speaker identification system.

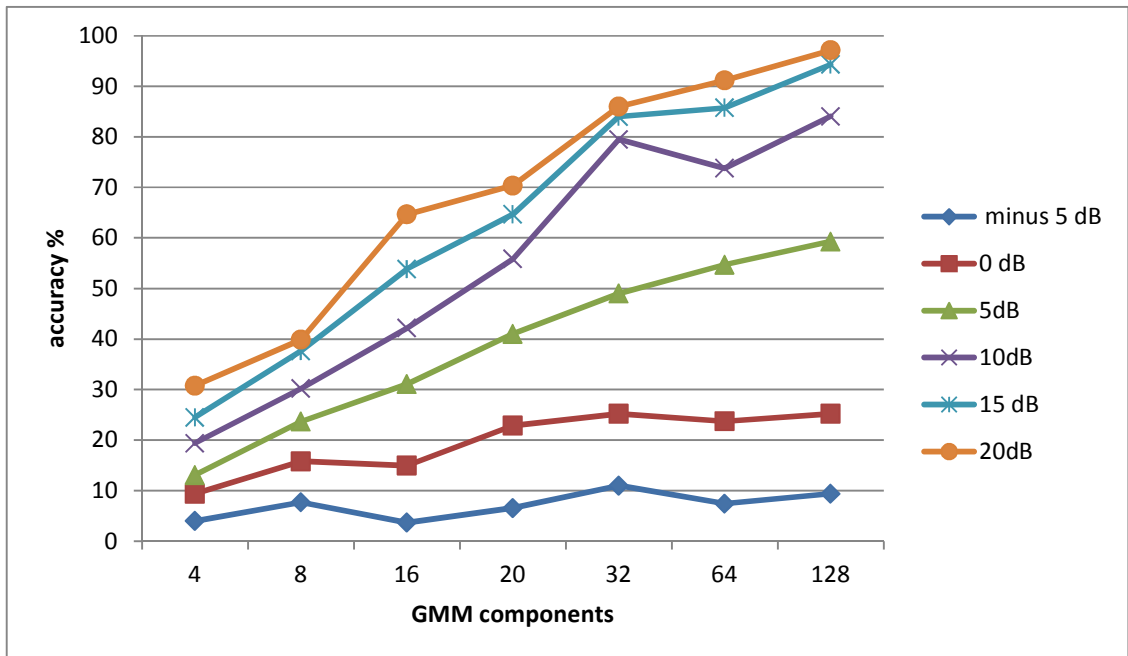


Figure 4.5: effect of changing GMM components on system accuracy for different levels of noise

Unlike ENV-based method, TFS-based method causes the accuracy of system to increase steadily when the number of GMM components has been increased. The highest accuracy has been achieved for the highest number of component, which is $K = 128$ in this case. This behavior has been observed for all SNRs.

Chapter 5. CONCLUSION

The goal of this project was to develop a neural response based speaker identification system, which takes into account the processing strategies employed by the auditory system. Usually the traditional techniques like MFCC, LPC, and PLP are implemented using the feature directly taken from the acoustic signal. But in this project, the neural response based approach used a physiologically-based model of the auditory system which simulated the responses of the AN fibers in the peripheral auditory system. The input to the model was the acoustic signals from the speakers, and the responses of the model were simulated for a wide range of characteristic frequencies. Two types of features, ENV and TFS, were extracted from the AN model output, which was used as an input to the classifier. Two types of classifiers, GMM and HMM, were employed in this project to train and test the proposed speaker identification system.

The results obtained from this project revealed that the performance of the proposed neural response based system was better than the performance of the traditional acoustic feature (MFCC) based speaker identification system, especially under noisy conditions. The result also showed that TFS response based identification system performed better than ENV response based system, meaning that the TFS contains more information related to the identity of speakers than in the ENV information. This could be related to the phase-locking properties of the AN fibers to the individual stimulus frequency. It is well-known that neural responses show phase-locking to a periodic input up to a frequency range of ~ 4 kHz at the level of auditory nerve. So, the spikes occur with a fixed delay or its multiples, and thus a distinct peak appears in the inter-spike interval

histogram. However, when noise is added to the periodic input signal, the spike interval histogram still shows a peak around the same interval, meaning that the neural responses are very robust to noise.

In this project, the performance of the proposed system using GMM and HMM as a classifier was also evaluated. The preliminary result showed that GMM performed better than using HMM with the neural responses from the peripheral auditory system. However, the parameters for HMM were not optimized in this case.

Finally, the effect of parameters on the accuracy of the identification was evaluated for a range of GMM components K , from 4 to 128. The performance of the TFS-based method showed an approximately linear increase in accuracy as a function of number of GMM components, and this trend was found to be true for almost all SNRs. However, ENV-based method showed the best accuracy for a range component numbers from ~20 to 32.

5.1. *Limitations*

The main limitation of the proposed method was that it was computationally very expensive compared to the traditional acoustic feature based identification systems, because the neural responses to the acoustic signal were simulated for a wide range of characteristic frequencies. Among the neural response based systems, TFS was more time consuming than the ENV based system.

5.2. *Future study*

In this study, the number of characteristic frequency used in the proposed method was 32, and the effect of increasing CFs on the accuracy was not evaluated. As a future

study, the effect of this number on the performance of the proposed system could be evaluated for both GMM and HMM as classifiers.

The accuracy of the neural response based system was evaluated using only white Gaussian noise. However, this method could be extended to evaluate performance using other types of noise, such as car, train, street, cocktail party, etc.

The parameters for HMM as a classifier were not optimized in this study, which could be done as a future study for both neural response and MFCC based speaker identification system.

REFERENCES

- Abuku, M., Azetsu, T., Uchino, E., & Suetake, N. (2010). *Application of peripheral auditory model to speaker identification*. Paper presented at the Nature and Biologically Inspired Computing (NaBIC), 2010 Second World Congress on.
- Atal, B. S. (2005). Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification. *The Journal of the Acoustical Society of America*, 55(6), 1304-1312.
- Azetsu, T., Abuku, M., Suetake, N., & Uchin, E. (2012). *Speaker Identification in Noisy Environment with Use of the Precise Model of the Human Auditory System*. Paper presented at the Proceeding of the international multi Conference of Engineers and Computer Scientists Vol1, IMECS, Hong Kong.
- Baker, J. (1975). The DRAGON system--An overview. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 23(1), 24-29.
- Bruce, I. C., Sachs, M. B., & Young, E. D. (2003). An auditory-periphery model of the effects of acoustic trauma on auditory nerve responses. *The Journal of the Acoustical Society of America*, 113(1), 369-388.
- Deng, L., & Geisler, C. D. (1987). A composite auditory model for processing speech sounds. *The Journal of the Acoustical Society of America*, 82(6), 2001-2012.
- Flanagan, J. L. (1960). Models for approximating basilar membrane displacement. *The Bell System Technical Journal*.
- Furui, S. (1981). Cepstral analysis technique for automatic speaker verification. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 29(2), 254-272.
- Geisler, C. D. (1976). Mathematical models of the mechanics of the inner ear *Auditory System* (pp. 391-415): Springer.
- Hall, J. L. (1981). Observations on a nonlinear model for motion of the basilar membrane. *Hearing Research and Theory*, 1, 1-61.
- Hermansky, H., Kohn, P., Morgan, N., & Bayya, A. (1992). *RASTA-PLP speech analysis technique*. Paper presented at the Acoustics, Speech, and Signal Processing, 1992. ICASSP-92 Vol 1., 1992 IEEE International Conference on.
- Hewitt, M. J., & Meddis, R. (1993). Regularity of cochlear nucleus stellate cells: a computational modeling study. *The Journal of the Acoustical Society of America*, 93(6), 3390-3399.

- Hines, A., & Harte, N. (2012). Speech intelligibility prediction using a neurogram similarity index measure. *Speech Communication*, 54(2), 306-320.
- Li, Q., & Huang, Y. (2010). *Robust speaker identification using an auditory-based feature*. Paper presented at the Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on.
- Lyon, R. F. (1984). *Computational models of neural auditory processing*. Paper presented at the Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'84.
- Mammone, R. J., Zhang, X., & Ramachandran, R. P. (1996). Robust speaker recognition: A feature-based approach. *Signal Processing Magazine, IEEE*, 13(5), 58.
- Meddis, R. (1986). Simulation of mechanical to neural transduction in the auditory receptor. *The Journal of the Acoustical Society of America*, 79(3), 702-711.
- Patterson, R. D., Allerhand, M. H., & Giguere, C. (1995). Time-domain modeling of peripheral auditory processing: A modular architecture and a software platform. *The Journal of the Acoustical Society of America*, 98(4), 1890-1894.
- Pelecanos, J., & Sridharan, S. (2001). Feature warping for robust speaker verification.
- Pfeiffer, R., & Kim, D. (1973). Considerations of nonlinear response properties of single cochlear nerve fibers. *Basic Mechanisms in Hearing, Academic Press, New York*, 555-587.
- Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker verification using adapted Gaussian mixture models. *Digital signal processing*, 10(1), 19-41.
- Reynolds, D. A., & Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *Speech and Audio Processing, IEEE Transactions on*, 3(1), 72-83.
- Robert, A., & Eriksson, J. L. (1999). A composite model of the auditory periphery for simulating responses to complex sounds. *The Journal of the Acoustical Society of America*, 106(4), 1852-1864.
- Shao, Y., Srinivasan, S., & Wang, D. (2007). *Incorporating auditory feature uncertainties in robust speaker identification*. Paper presented at the Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on.
- Togneri, R., & Pullella, D. (2011). An overview of speaker identification: Accuracy and robustness issues. *Circuits and Systems Magazine, IEEE*, 11(2), 23-61.
- Wilson, B. S., Schatzer, R., Lopez-Poveda, E. A., Sun, X., Lawson, D. T., & Wolford, R. D. (2005). Two new directions in speech processor design for cochlear implants. *Ear and hearing*, 26(4), 73S-81S.

- Zhang, X., Heinz, M. G., Bruce, I. C., & Carney, L. H. (2001). A phenomenological model for the responses of auditory-nerve fibers: I. Nonlinear tuning with compression and suppression. *The Journal of the Acoustical Society of America*, 109(2), 648-670.
- Zilany, M. S., & Bruce, I. C. (2006). Modeling auditory-nerve responses for high sound pressure levels in the normal and impaired auditory periphery. *The Journal of the Acoustical Society of America*, 120(3), 1446-1466.
- Zilany, M. S., Bruce, I. C., Nelson, P. C., & Carney, L. H. (2009). A phenomenological model of the synapse between the inner hair cell and auditory nerve: long-term adaptation with power-law dynamics. *The Journal of the Acoustical Society of America*, 126(5), 2390-2412.